



International Journal of Research in Agronomy

E-ISSN: 2618-0618
P-ISSN: 2618-060X
© Agronomy
NAAS Rating (2026): 5.20
www.agronomyjournals.com
2026; SP-9(3): 130-135
Received: 16-12-2025
Accepted: 20-01-2026

Richa Dhurandher
Indira Gandhi Krishi
Vishwavidyalaya (IGKV), Raipur,
Chhattisgarh, India

Adarsh Kumar
Indira Gandhi Krishi
Vishwavidyalaya (IGKV), Raipur,
Chhattisgarh, India

Estimating Yield and Production of Paddy and Maize Crops of Dantewada district of Chhattisgarh using regression analysis

Richa Dhurandher and Adarsh Kumar

DOI: <https://www.doi.org/10.33545/2618060X.2026.v9.i3Sb.5049>

Abstract

This study estimates the yield and production of paddy and maize in Dantewada district, Chhattisgarh, using 10 years of data (2010-11 to 2019-20). Key variables analyzed include area, yield, production, irrigation, NPK consumption, rainfall, humidity, and temperature. ARIMA parameters indicated no auto-correlation for paddy, leading to the use of a linear regression model, while maize data showed auto-correlation, necessitating generalized least squares modeling. Forecasting for crop yield was extended to 2025 for both crops.

Keywords: ARIMA (P,D,Q), Linear models, Generalised linear models, Multi-collinearity

Introduction

Regression analysis is a method for assessing the linear relationship among variables. Multiple regression involves one dependent variable influenced by several independent variables. Studies cited include Wei and Molin (2020) ^[37], who estimated soybean yield using linear regression, Nazir *et al.* (2021) ^[21], who forecasted rice yield with a phenology-based algorithm and regression model, and Sellam and Poovammal (2016) ^[27] as well as Shastry (2017) ^[28], both of whom focused on crop yield predictions through regression analysis.

In time series analysis, residuals may show serial correlation or heteroscedasticity, contravening classical regression assumptions. A solution is the Generalized Least Squares (GLS) method, which maximizes likelihood estimation in regression models with ARIMA errors, as explained by Otto *et al.* (1987). Studies by Kadanali *et al.* (2019) focused on ARIMA model forecasting for wheat production, while Dritsakis *et al.* (2019) provided insights into time series analysis with ARIMA. Additionally, Kwasi *et al.* compared the forecasting abilities of multivariate VAR and univariate ARIMA models. In regression analysis, multicollinearity among dependent variables can harm model accuracy, leading to the elimination of high multi-collinear variables via correlation matrices. The identification of multicollinearity is detailed by Shrestha (2020) ^[30], while research conducted by Haitovsky (1969) ^[17] and Daoud (2017) ^[10] further explores its implications. Kim (2019) ^[18] elucidates the relationship between multicollinearity and misleading outcomes, and Schroeder *et al.* examine methods for diagnosing and treating this issue. Regression and ARIMA models can be integrated to enhance forecasting accuracy by leveraging their strengths, especially when data exhibits both linear correlations and temporal dependencies. Shumway *et al.* (2017) ^[31] discussed time series analysis and ARIMA models, while Mondal *et al.* (2014) investigated ARIMA's efficacy in predicting stock prices. Benvenuto *et al.* (2020) ^[9] analyzed the ARIMA model's application using COVID-19 pandemic data.

Material and Methods

Study area

This study has been conducted in Dantewada district of Chhattisgarh for the crops paddy and maize. It is located on the latitude of 18° 89' 0" N and longitude of 81° 34' 0" E. It is bounded

Corresponding Author:
Richa Dhurandher
Indira Gandhi Krishi
Vishwavidyalaya (IGKV), Raipur,
Chhattisgarh, India

by Bastar and Narayanpur to the north, Bijapur to the west and Sukma to the south and east with Kondagaon.

Data Description

Secondary data forms the basis of this investigation, which focuses on the region of Chhattisgarh, created from the bifurcation of Madhya Pradesh in 2000. Since then, new districts have emerged from the original districts multiple times. The study considers various variables, including area, gross irrigation area, NPK consumption, temperature ranges, humidity, and rainfall, along with yield and production. Outliers were removed to ensure model accuracy. The study employs the ARIMA (p, d, q) model to assess serial correlation in time series data for paddy, maize, and jowar crops in Dantewada, selecting significant predictor variables through a correlation matrix, and utilizing linear and generalized linear models as needed.

Linear Regression

Linear regression is a statistical method that models the relationship between a dependent variable y and one or more independent variables (X_1, X_2, X_n) assuming a linear relationship where changes in the independent variable(s) correspond to constant changes in the dependent variable.

$$Y = \beta_0 + \beta_1 \times X_1 + \beta_2 \times X_2 \dots \beta_n \times X_n + \epsilon \quad \dots(1)$$

Where Y is the dependent variable, $X_1, X_2, \dots, X_n$ are the independent variables, β_0 is the intercept, $\beta_1, \beta_2, \dots, \beta_n$ are the coefficients (slopes) of the independent variables, ϵ is the error term, which represents the difference between the observed and predicted values of Y .

Generalised least square model

Generalized least squares (GLS) is a method used to estimate unknown parameters in linear regression models when the standard assumptions regarding residuals—such as zero mean, normality, or equal variances—are violated. It addresses these issues by incorporating a weighting matrix that assigns varying weights to data points according to their estimated variances.

- **Estimate the error covariance matrix:** This matrix captures how the errors are related to each other. It might show higher variances for certain groups of data points.
- **Apply the weights:** Each data point in the regression equation is multiplied by a weight based on the error covariance matrix. This gives more influence to data points

with lower error variance and less influence to those with higher variance.

- **Minimize the weighted squared residuals:** Similar to OLS, GLS minimizes the sum of squared residuals, but these residuals are now weighted based on the error covariance matrix.

Multi-collinearity

Multicollinearity occurs in regression analysis when two or more independent predictor variables are strongly correlated. This correlation complicates the evaluation of each variable's individual effect on the dependent variable, making it challenging to identify their true influence. As a result, the model's predictions may become less accurate due to difficulties in assessing the underlying impacts of these correlated variables.

Auto Regressive Integrated Moving Average (ARIMA)

ARIMA, which stands for Auto Regressive Integrated Moving Average, is a statistical method prevalent in time series forecasting and analysis. It effectively captures various temporal patterns, making it applicable in fields such as economics, finance, epidemiology, and weather forecasting.

The general ARIMA (p, d, q) model can be expressed with the following equation:

$$Y_t = \mu + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \dots + \phi_p Y_{t-p} + \epsilon_t - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2} - \dots - \theta_q \epsilon_{t-q} \quad \dots(2)$$

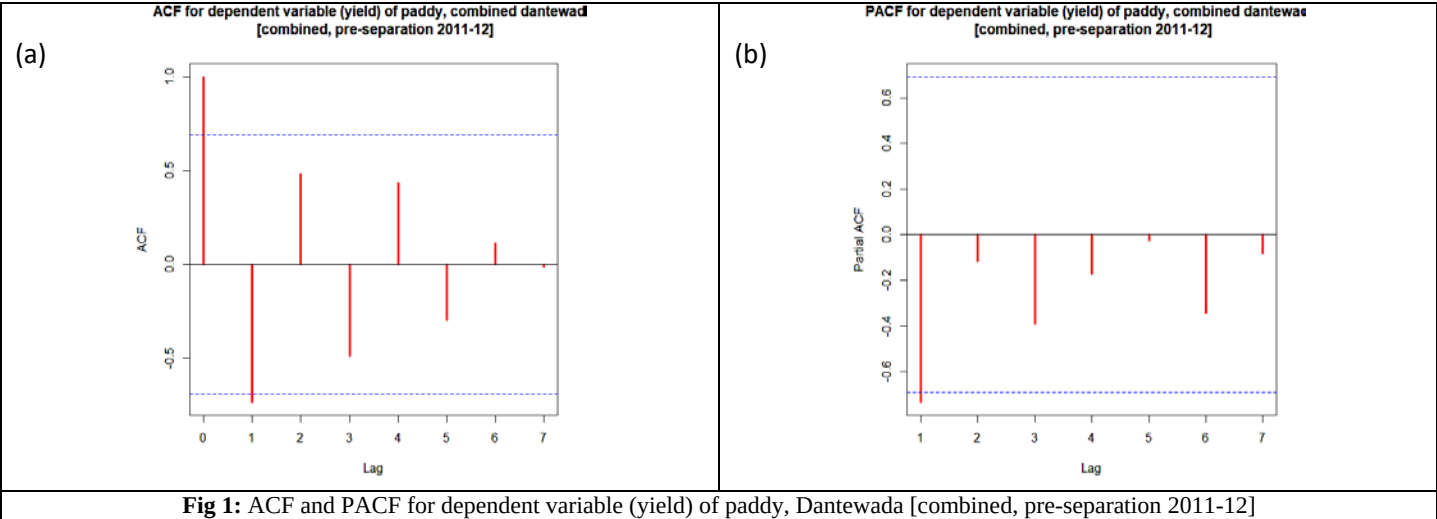
Where, Y_t is the actual value at time t , μ is the constant mean (optional, may not be present in all models), ϕ_1 (phi) are the autoregressive parameters (ϕ_1 to ϕ_p). These represent the coefficients of the past p values of Y_t , Y_{t-1} to Y_{t-p} are the lagged values of Y_t , influencing the current value (Y_t), ϵ_t is the white noise error term at time t (represents unpredictable random shocks), θ (theta) are the moving average parameters (θ_1 to θ_q). These represent the coefficients of the past q forecast errors (ϵ_{t-1} to ϵ_{t-q}).

Model evaluation

After putting all of the estimate methods through diagnostic plots and goodness of fit metrics like R^2 , Adj R^2 , and their P-values, the process is concluded.

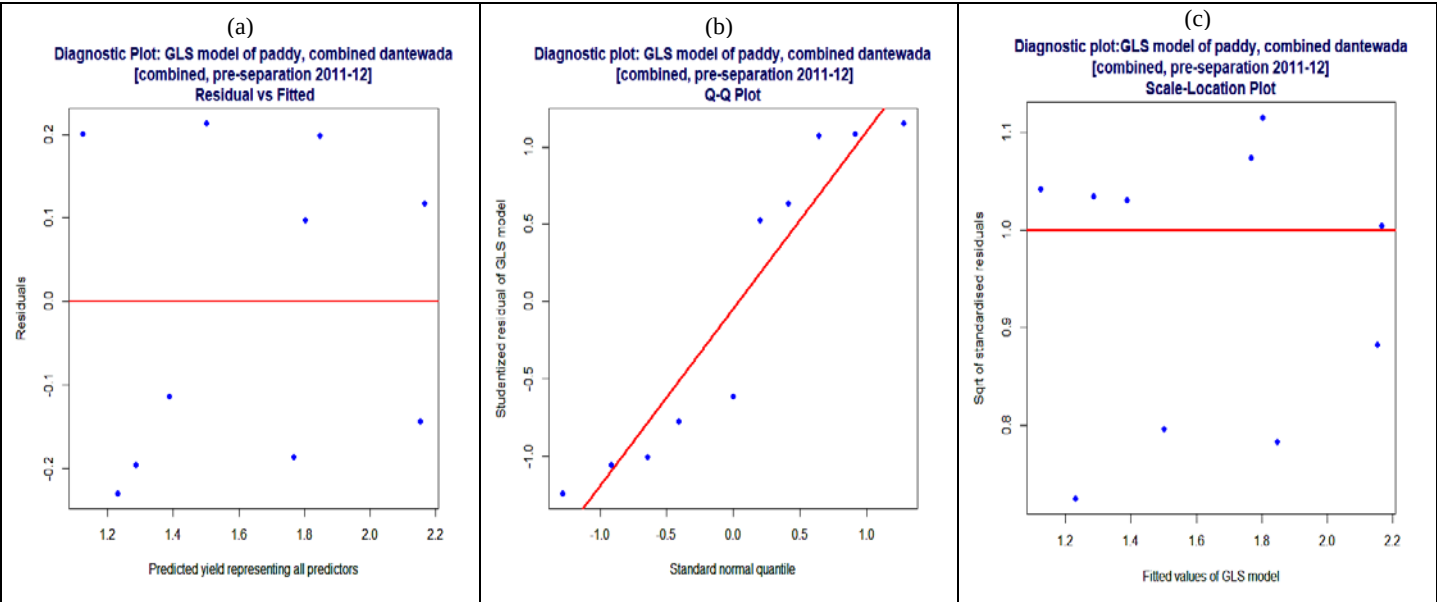
Results and Discussions

Paddy crop of Dantewada



From the time series data of the yield of the paddy crop in Dantewada, auto-regressive integrated moving average, ARIMA (p, d, q) are determined to confirm the presence or absence of auto-correlation.

Using acf() and pacf() functions of R, p=1 and q=1 as the correlations are beyond the significant bounds after differencing twice to ensure stationarity. Hence ARIMA (1, 2, 1) was obtained.



The diagnostic plots indicating a good fitting model with $R^2 = 0.7823$ and $\text{Adj } R^2 = 0.4921$ and the finalized linear statistical model is given below.

Further, the yield, predicted yield, standard error and confidence interval of the finalized model are given in Table 1.

$$y = 220 + 0.00870 \times x_3 - 8.22 \times x_4 + 6.36 \times x_5 - 1.47 \times x_6 \quad \dots (3)$$

Table 1: Comparison of yield observed with the predicted one along with confidence interval for paddy, Dantewada

Year	Yield	Predicted Yield	Standard Error	Lower (95% CI)	Upper (95% CI)
2010-11	1.900	1.80261	0.1722679	0.761466	2.843754
2011-12	1.275	1.388719	0.1722679	0.347575	2.429863
2012-13	2.285	2.167586	0.1722679	1.126442	3.208730
2013-14	2.010	2.154304	0.1722679	1.113160	3.195448
2014-15	2.045	1.846885	0.1722679	0.805740	2.888029
2015-16	1.000	1.230246	0.1722679	0.189102	2.271390
2016-17	1.715	1.501379	0.1722679	0.460235	2.542523
2017-18	1.090	1.286522	0.1722679	0.245378	2.327666
2018-19	1.325	1.123946	0.1722679	0.082802	2.165090
2019-20	1.580	1.766689	0.1722679	0.725545	2.807833

Moreover, goodness of fit plot between predicted yield and yield is further done and depicted in Figure 3.

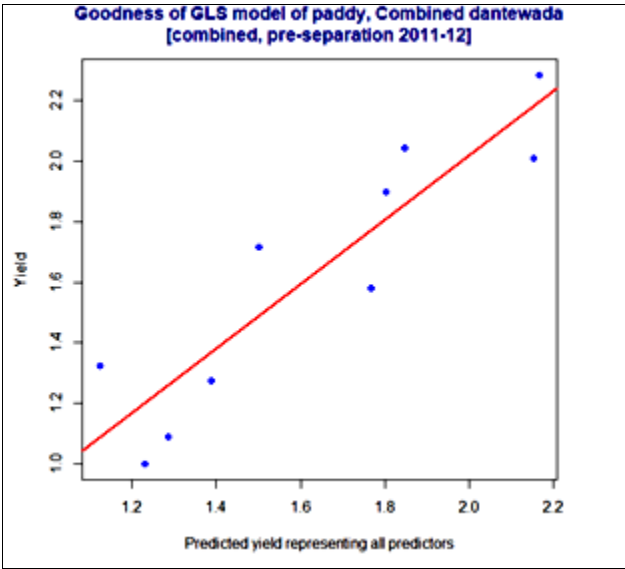


Fig 3: Goodness of the prediction model in terms of predicted vs observed yield for paddy, Dantewada

After fitting and finalizing the model, a forecast for the yield and production has been made upto 2025.

Table 2: Forecast of yield and production for 2020-21 to 2024-25 based on the projected area under paddy, Dantewada

Year	Projected area	Predicted yield	Predicted production
2020	160626	1.321641	212289.907
2021	163004.3	1.266141	206386.427
2022	165382.5	1.210642	200219.001
2023	167760.8	1.155142	193787.546
2024	170139	1.099643	187092.16

Maize crop of Dantewada

From the time series data of the yield of the maize crop in Dantewada, auto-regressive integrated moving average, ARIMA (p, d, q) are determined to confirm the presence or absence of

auto-correlation. Using acf() and pacf() functions of R, p=1 and q=1 as the correlations are beyond the significant bounds after differencing twice to ensure stationarity. Hence ARIMA (1, 2, 1) was obtained.

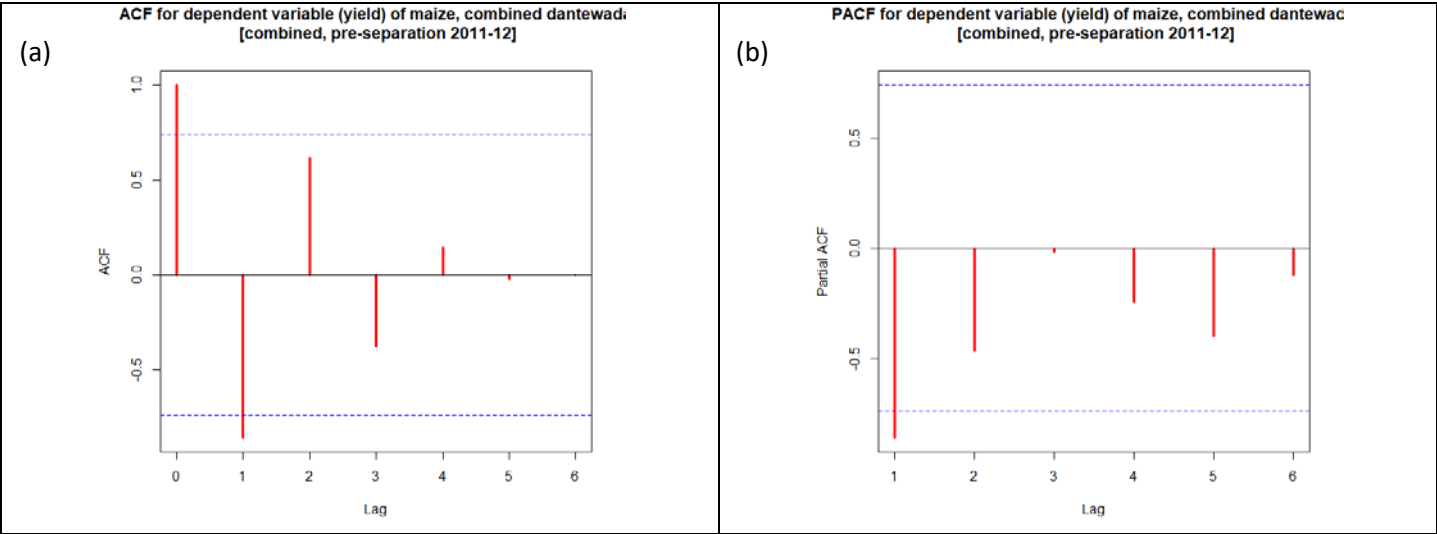


Fig 4: ACF and PACF for dependent variable (yield) of maize, Dantewada [combined, pre-separation 2011-12]

The values of p=1 and q=1 from Step-1 indicating the presence of serial correlation. As a result, generalised least square is used instead of linear model. Further, to know the order of predictor variables for entering into the generalised least square model, correlation matrix was used. From the correlation matrix, the variables having highest correlation with dependent variable

were selected in descending order. The orderof the variables is as follows: ‘area.h’, ‘X20.per.gia.prop’. The generalized linear model for the maize crop is given below:

$$y = 3.549 - 0.000488 \times x_1 + 14.686 \times x_3 + x_2$$

...(4)

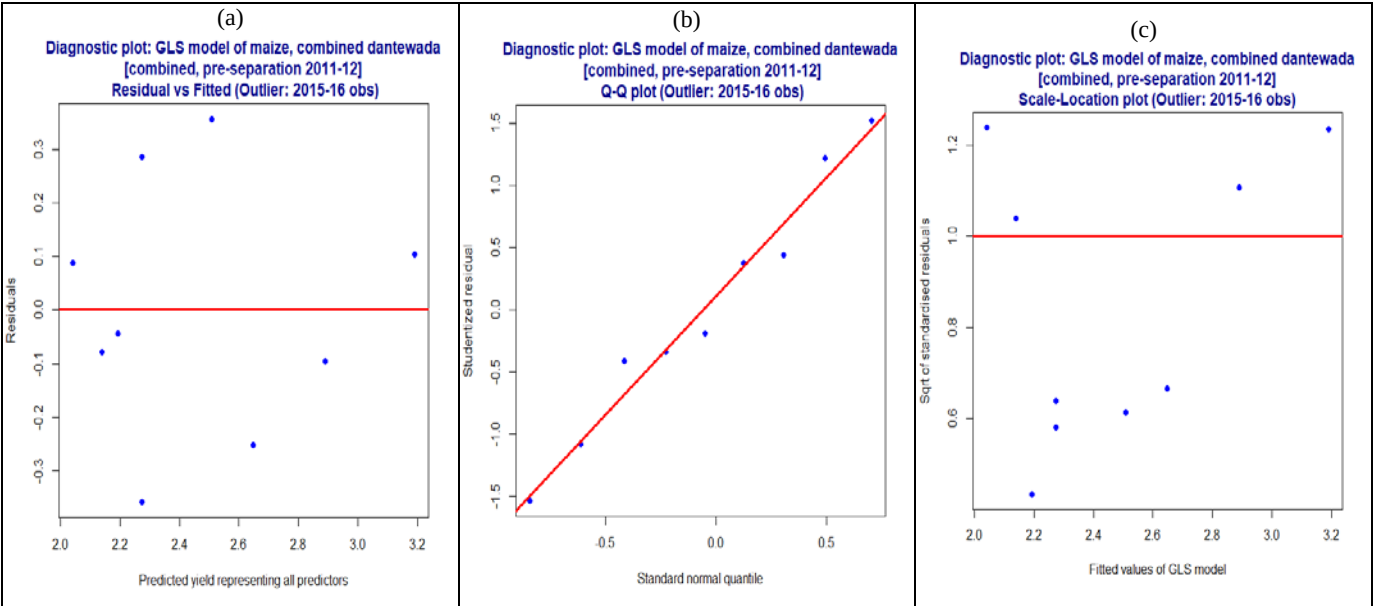


Fig 5: Three major diagnostic plots for the linear model of maize, Dantewada

Since the diagnostic plots were seemed to fit well, the predicted yield, standard error and confidence interval of the generalised least square model was finalized. Further, the yield, finalized model are given in Table 4.

Table 3: Comparison of yield observed with the predicted one along with confidence interval for maize, Dantewada

Year	Yield	Predicted yield	Standard error	Confidence interval 95%	
				Lower	Upper
2010-11	2.13	2.041695	0.2597	1.276107	2.807283
2011-12	2.06	2.138716	0.2597	1.373128	2.904304
2012-13	1.915	2.273684	0.2597	1.508097	3.039272
2013-14	2.56	2.273502	0.2597	1.507914	3.03909
2014-15	2.15	2.19403	0.2597	1.428443	2.959618
2016-17	2.865	2.508209	0.2597	1.742621	3.273797
2017-18	2.395	2.647911	0.2597	1.882323	3.413498
2018-19	2.795	2.89025	0.2597	2.124662	3.655837
2019-20	3.295	3.191332	0.2597	2.425744	3.95692

Moreover, goodness of fit plot between predicted yield and yield is further done and depicted in Fig.6. It is very clear that the data points are close to the fitted line indicating a good fit with

$R^2 = 0.7348$ (P-value: 0.003141) and Adj $R^2 = 0.6969$

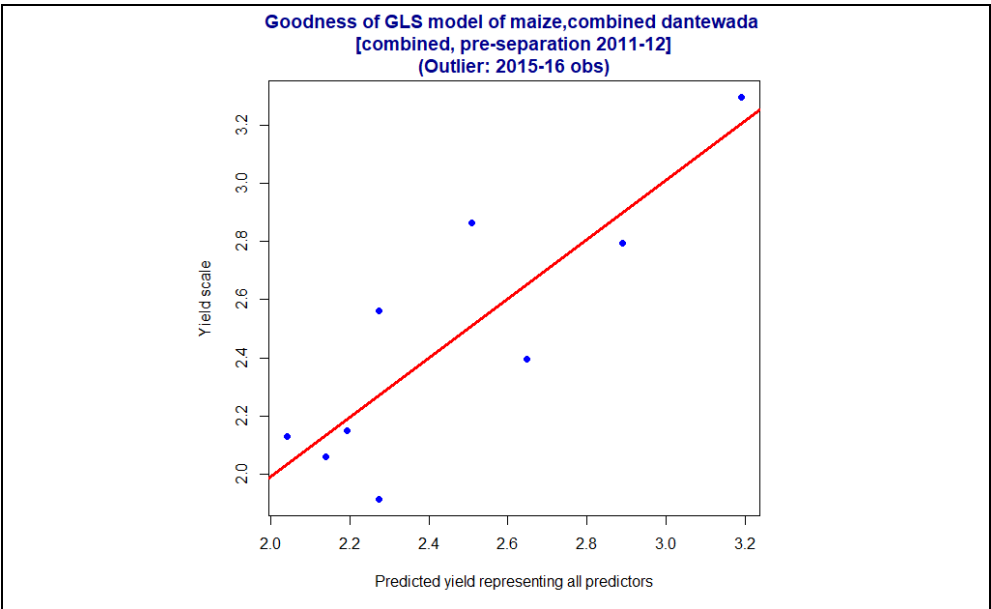


Fig 6: Goodness of the prediction model in terms of predicted vs observed yield for maize, Dantewada

A forecast upto 2025 has been made for yield and production after finalizing the model.

Table 4: Forecast of yield and production for 2020-21 to 2024-25 based on the projected area under maize, Dantewada

Year	Projected area	Predicted yield	Predicted production
2020	4514.533	53.00856	239308.9
2021	4322.776	53.83342	232709.8
2022	4131.018	54.65827	225794.3
2023	3939.261	55.48312	218562.5
2024	3747.503	56.30798	211014.3

Conclusion

In this study, predicted yield is very much close to the actual yield indicating that the finalized model is a good fit. Through forecasting of yield of paddy and maize upto 2025, there is showing an increasing yield in future which is a good index.

References

- Anonymous. List of districts of Chhattisgarh. Wikipedia. 2023. Available from: https://en.wikipedia.org/wiki/List_of_districts_of_Chhattisgarh
- Arunachalam R, Balakrishnan V. Statistical modeling for wheat (*Triticum aestivum*) crop production. *Int J Stat Appl*. 2012;2(4):40-46.
- Bahrami M, Shabani A, Mahmoudi MR, Didari S. Determination of effective weather parameters on rainfed wheat yield using backward multiple linear regressions based on relative importance metrics. *Complexity*. 2020;2020:1-10.
- Barnett V, Lewis T. Outliers in statistical data. New York: Wiley; 1994.
- Bartlett MS. The use of transformations. *Biometrics*. 1947;3(1):39-52.
- Belov AG. A mathematical-statistics approach to the least squares method. *Comput Math Model*. 2018;29:30-41.
- Chatfield C. Exploratory data analysis. *Eur J Oper Res*. 1986;23(1):5-13.
- Chaudhary JL. Rainfall-rice yield relationship: a case study for Chhattisgarh region of Madhya Pradesh in central India. 1999.
- Ciotti M, Angeletti S, Minieri M, Giovannetti M, Benvenuto D, Pascarella S, *et al*. COVID-19 outbreak: an overview. *Chemotherapy*. 2020;64(5-6):215-223.
- Daoud JI. Multicollinearity and regression analysis. *J Phys Conf Ser*. 2017;949(1):012009.
- Directorate of Economics and Statistics. Area, production and productivity of various districts of Chhattisgarh for the years 2010-11 to 2019-20. Govt. of Chhattisgarh; 2023. Available from: http://aps.dac.gov.in/APY/Public_Report1.aspx
- Diwan UK, Puranik HV, Das GK, Chaudhary JL. Yield prediction of wheat at preharvest stage using regression based statistical model for 8 districts of Chhattisgarh, India. *Int J Curr Microbiol Appl Sci*. 2018;7(1):2180-2183.
- Draper NR, Smith H. Applied regression analysis. New York: John Wiley & Sons; 1998.
- Ekanayake EMP, Wickramasinghe LCD, Weliwatta RT. Use of regression techniques for rice yield estimation in the North-Western province of Sri Lanka. *Ceylon J Sci*. 2021;50(4):439-447.
- Ferdushi KF, Hossain MK, Kamil AA. Production risk with feasible generalized least square. *J Phys Conf Ser*. 2020;1641(1):012109.
- Ferguson IS, Leech JW. Generalized least squares estimation of yield functions. *For Sci*. 1978;24(1):27-42.
- Haitovsky Y. Multicollinearity in regression analysis: comment. *Rev Econ Stat*. 1969;486-489.
- Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, *et al*. PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res*. 2019;47(D1):D1102-D1109.
- Montgomery DC, Peck EA, Vining GG. Introduction to linear regression analysis. New York: John Wiley & Sons; 2021.
- NASA. Prediction of worldwide energy resources: The power project. 2023. Available from: <https://power.larc.nasa.gov/>
- Nazir A, Ullah S, Saqib ZA, Abbas A, Ali A, Iqbal MS, *et al*. Estimation and forecasting of rice yield using phenology-based algorithm and linear regression model on Sentinel-II satellite data. *Agriculture*. 2021;11(10):1026.
- R Core Team. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2023.
- Rajaratnam A, Parmar RS, Vaishnav PR. Estimating models for area, production and productivity trends of tobacco (*Nicotiana tabacum*) crop for Anand region of Gujarat State. *Indian J Appl Sci*. 2010;10:2419-2425.
- Ran Y, Chen H, Ruan D, Liu H, Wang S, Tang S, *et al*. Identification of factors affecting paddy yield gap in Southwest China: an experimental study. *PLoS One*. 2018;13(11).
- Rossi R, Murari A, Gaudio P, Gelfusa M. Upgrading model selection criteria with goodness of fit tests for practical applications. *Entropy (Basel)*. 2020;22(4):447.
- Sahu G, Nishad D, Lakhera ML, Mishra P, Joshi RP. Influence of area and yield on the production of maize in Chhattisgarh plain. *J Pharmacogn Phytochem*. 2018;1:71-75.
- Sellam V, Poovammal E. Prediction of crop yield using regression analysis. *Indian J Sci Technol*. 2016;9(38):1-5.
- Shastri A, Sanjay HA, Bhanusree E. Prediction of crop yield using regression techniques. *Int J Soft Comput*. 2017;12(2):96-102.
- Shirazi SM, Yusop Z, Zardari NH, Ismail Z. Effect of irrigation regimes and nitrogen levels on the growth and yield of wheat. *Adv Agric*. 2014.
- Shrestha N. Detecting multicollinearity in regression analysis. *Am J Appl Math Stat*. 2020;8(2):39-42.
- Shumway R, Stoffer D. Time series: a data analysis approach using R. Boca Raton: Chapman and Hall/CRC; 2019.
- Singh DP, Kumar D, Paikra MS, Kusro PS. Developing statistical models to study the growth rates of paddy crops in different districts of Chhattisgarh. *Am Int J Res Formal Appl Nat Sci*. 2014;5(1):102-104.
- Singh R, Brahme R, Singh VB. Growth in area, production and productivity of kharif paddy in Chhattisgarh. *Int J Agric Sci Res*. 2018;8(4):65-72.
- Usman A, Tukur K, Suleiman A, Abdulkadir A, Ibrahim H. The use of the weighted least squares method when the error variance is heteroscedastic. *Benin J Stat*. 2019;2:85-93.
- Uyanık GK, Güler N. A study on multiple linear regression analysis. *Procedia Soc Behav Sci*. 2013;106:234-240.
- Wasnik S, Pandey S, Patel P, Choudary V. Trends in area, production and yield of paddy, wheat and gram in Chhattisgarh State: a critical analysis. *Pharma Innov J*. 2022;11(6):1397-1402.
- Wei MCF, Molin JP. Soybean yield estimation and its components: a linear regression approach. *Agriculture*. 2020;10(8):348.